

Finding Community Structure with Performance Guarantees in Complex Networks

Thang N. Dinh and My T. Thai
Computer & Information Science & Engineering
University of Florida, Gainesville, FL, 32611,
Email: {tdinh,mythai}@cise.ufl.edu

Abstract—Many networks including social networks, computer networks, and biological networks are found to divide naturally into communities of densely connected individuals. Finding community structure is one of fundamental problems in network science. Since Newman’s suggestion of using *modularity* as a measure to qualify the goodness of community structures, many efficient methods to maximize modularity have been proposed but without a guarantee of optimality. In this paper, we propose two polynomial-time algorithms to the modularity maximization problem with theoretical performance guarantees. The first algorithm comes with a *priori guarantee* that the modularity of found community structure is within a constant factor of the optimal modularity when the network has the power-law degree distribution. Despite being mainly of theoretical interest, to our best knowledge, this is the first approximation algorithm for finding community structure in networks. In our second algorithm, we propose a *sparse metric*, a substantially faster linear programming method for maximizing modularity and apply a rounding technique based on this sparse metric with a *posteriori approximation guarantee*. Our experiments show that the rounding algorithm returns the optimal solutions in most cases and are very scalable, that is, it can run on a network of a few thousand nodes whereas the LP solution in the literature only ran on a network of at most 235 nodes.

I. INTRODUCTION

Many complex systems of interest such as the Internet, social, and biological relations, can be represented as networks consisting a set of *nodes* which are connected by *edges* between them. Research in a number of academic fields has uncovered unexpected structural properties of complex networks including small-world phenomenon [1], power-law degree distribution [2], and the existence of community structure [3] where nodes are naturally clustered into tightly connected modules, also known as communities, with only sparser connections between them.

The detection of community structures in networks is an important problem that has drawn an enormous amount of research effort [4]. A huge benefit of identifying community structure is that one can infer semantic attributes for different communities. For example in social networks, the attributes for a community can be common interest or location, and for metabolic networks the attribute could be a common function. Moreover, the relative independence among different communities allows the examining of each community individually, and an analysis of network at a higher-level of structure.

There are a wide variety of definitions for communities. In general, definitions can be classified into two main categories:

local definitions and *global definitions*. In local definitions, only the group of nodes and its immediate neighborhood are considered, ignoring the rest of the network. For example, communities can be defined as maximal *cliques*, *quasi-cliques*, *k-plexes*. The most famous definitions in this category are notions of *strong community*, where each node has more neighbors inside than outside the community, and *weak community*, where the total number of inner edges must be at least half of the number of outgoing edges.

In global definitions, communities can be only recognized by analyzing the network as a whole. This type of definitions is especially suitable when the next phase after the community detection is to optimize a global quantity, for example, minimizing the inter-group communication cost. The most widely-used quantity function in the global category is Newman’s *modularity* which is defined as the number of edges falling within communities minuses the expected number in an equivalent network with edges placed at random [5]. A higher value of modularity, a better community structure. Thus, identifying a good community structure of a given network becomes finding a partition of networks so as to maximize the modularity of this partition, called modularity maximization problem.

Since the introduction of modularity, maximizing modularity has become primal approaches to detect community structure. Numerous computational methods have been proposed, based on agglomerative hierarchical clustering [6], simulated annealing [7], genetic search [8], extremal optimization [9], spectral clustering [10], multilevel partitioning [11], and many others. For a comprehensive view of community detection methods, we refer to an excellent survey of S. Fortunato and C. Castellano [4].

Unfortunately, Brandes et al. [12] have shown that modularity maximization is an NP-hard problem, thereby denying the existence of polynomial-time algorithms to find optimal solutions. Thus, it is desirable to design polynomial-time approximation algorithms to find partitioning with a theoretical performance guarantee on the modularity values.

In contrary to the vast amount of work on maximizing modularity, the only known polynomial-time approach to find a good community structure with guarantees is due to G. Agarwal and D. Kempe [13] in which they rounded the fractional solution of a linear programming (LP). The value obtained by the LP is an upper bound on the maximum achievable

modularity. Thus, their approach provide a posteriori guarantee on the error bound. In fact, the modularity values found by their approach are optimal for many network instances comparing with the optimal modularity values provided by expensive exact algorithms in [14]. The main drawback of the approach is the large LP formulation that consumes both time and memory resources. As shown in their paper, the approach can only be used on the networks of up to 235 nodes. Secondly, while the approach performs well on all considered networks, it does not promise any priori guarantees as provided by *approximation algorithms*.

In this paper, we address the main drawback of the rounding LP approach by introducing an improved formulation, called *sparse metric*. We show that our new technique substantially reduces the time and memory requirements both theoretically and experimentally without any trade-off on the quality of the solution. The size of solved network instances raises from hundred to several thousand nodes while the running time on the medium-instances are sped up from 10 to 150 times.

Our second contribution is an approximation algorithm that finds a community structure in networks with modularity values within a constant factor of the optimum when the considered networks have power-law degree distributions. To our best knowledge, it is the first approximation algorithm for finding community structure in networks. The algorithm is not only of theoretical interest, but also establish a connection between the power-law degree distribution properties and the presence of community structure in complex networks. Since community structure are often observed together with the power-law property, studying the community structure detection under power-law network models is of great important.

Organization. We present definitions and notions in Section II. We propose in Section III the *sparse metric* technique to efficiently maximize modularity via rounding a linear programming. An approximation algorithm for networks with the power-law degree distribution (so-called power-law networks) is introduced in Section IV. We show experimental results for the *sparse metric* in Section V to illustrate the time efficiency over the previous approach. Finally, in Section VI we summarize our results and discuss on limitation of modularity as well as the corresponding resolution.

II. PRELIMINARIES

A network can be represented as an undirected graph $G = (V, E)$ consisting of $n = |V|$ nodes and $m = |E|$ edges. The adjacency matrix of G is denoted by $A = (A_{i,j})$, where $A_{i,j} = A_{j,i} = 1$ if i and j share an edge and $A_{i,j} = A_{j,i} = 0$ otherwise.

A modularity maximization problem asks us to identify a community structure $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$ of a given graph where each disjoint subsets C_i are called *communities* and $\bigcup_{i=1}^k C_i = V$ so as to maximize the modularity of \mathcal{C} . Note that k is not a pre-defined value. The *modularity* [10] of \mathcal{C} is the fraction of the edges that fall within the given communities minus the expected number of such fraction if edges were distributed at random. The randomization of the edges is done

so as to preserve the degree of each vertex. If nodes i and j have degrees d_i and d_j , then the expected number of edges falling between i and j is $\frac{d_i d_j}{2m}$. Thus, the modularity, denoted Q , is then

$$Q(\mathcal{C}) = \frac{1}{2m} \sum_{i,j} (A_{i,j} - \frac{d_i d_j}{2m}) \delta_{ij} \quad (1)$$

where $\delta_{ij} = \begin{cases} 1, & \text{if } i, j \text{ are in the same communities} \\ 0, & \text{otherwise.} \end{cases}$

We also define modularity matrix B [10] as

$$B_{ij} = A_{ij} - \frac{d_i d_j}{2m}.$$

We note that each row and column of B sum up to zero, hence, B always has the vector $(1, 1, 1, \dots)$ as one of its eigenvectors. The same property is also known for the network Laplacian matrix $L = D - A$, where D is diagonal matrix with the i th entry to be d_i . Laplacian matrix L is widely-used in spectral methods for the graph partitioning that is closely related to our community detection problem. We note that the major difference between the modularity matrix and the Laplacian matrix is that L is positive-definite while B is indefinite. As a consequence, while approximation algorithms for the graph partitioning problem using Laplacian matrix L are available, it is not known if such algorithms are possible for the modularity maximization problem.

III. LINEAR PROGRAMMING BASED ALGORITHM

A. The Linear Program and The Rounding

The modularity maximization problem can be formulated as an Integer Linear Programming (ILP). The linear program has one variable $d_{i,j}$ for each pair (i, j) of vertices to represent the “distance” between i and j i.e.

$$d_{i,j} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ are in the same community} \\ 1 & \text{otherwise.} \end{cases}$$

In other words, $d_{i,j}$ is equivalent to $1 - \delta_{i,j}$ in the definition (1) of modularity. Thus, the objective function to be maximized can be written as $\sum_{i,j} B_{i,j} (1 - d_{i,j})$. We note that there should be no confusion between $d_{i,j}$ the variable representing the distance between vertices i and j and constant d_i (or d_j), the degree of node i (or j). The ILP to maximize modularity ($\text{IP}_{\text{complete}}$) is as follows

$$\text{maximize} \quad \frac{1}{2m} \sum_{i,j} B_{i,j} (1 - d_{i,j}) \quad (2)$$

$$\text{subject to} \quad d_{i,j} + d_{j,k} - d_{i,k} \geq 0, \quad \forall i < j < k \quad (3)$$

$$d_{i,j} - d_{j,k} + d_{i,k} \geq 0, \quad \forall i < j < k \quad (4)$$

$$-d_{i,j} + d_{j,k} + d_{i,k} \geq 0, \quad \forall i < j < k \quad (5)$$

$$d_{i,j} \in [0, 1], \quad i, j \in [1..n], \quad (6)$$

Constraints (3), (4), and (5) are well-known *triangle inequalities* that guarantee the values of $d_{i,j}$ are consistent to each

other. They imply the following transitivity: if i and j are in the same community and j and k are in the same community, then so are i and k . By definition, $d_{i,i} = 0 \forall i$ and can be removed from the ILP for simplification.

To avoid solving ILP, that is also NP-hard, we instead solve the LP relaxation of the ILP, obtained by replacing the constraints $d_{i,j} \in \{0, 1\}$ by $d_{i,j} \in [0, 1]$. We shall refer to the IP described above as $\text{IP}_{\text{complete}}$ and its relaxation as $\text{LP}_{\text{complete}}$. If the optimal solution of this relaxation is an integral solution, which is very often the case [14], we have a partition with the maximum modularity. Otherwise, we resort on rounding the fractional solution and use the value of the objective as an upper-bound that enables us to lower-bound the gap between the rounded solution and the optimal integral solution.

G. Agarwal and D. Kempe [13] use a simple rounding algorithm proposed by Charikar et al. [15] for the *correlation clustering* problem [16]. The values of $d_{i,j}$ are interpreted as a metric “distance” between vertices. The algorithm repeatedly groups all vertices that are close by to a vertex into a community. The final community structure are then refined by a Kernighan-Lin [17] based local search method.

Since the rounding phase is comparatively simple, the burden of both time and memory comes from solving the large LP relaxation. The LP has $\binom{n}{2}$ variables and $3\binom{n}{3} = \theta(n^3)$ constraints that is about half a million constraints for a network of 100 vertices, thereby limiting the size of networks to few hundred nodes. Thus, there is a need to achieve the same guarantees with smaller resource requirements. By combining mathematical approach with combinatorial techniques, we achieve this goal in next subsection.

B. The Sparse Metric

In this subsection, we devise an improved LP formulation for the modularity maximization problem with much fewer number of constraints while getting the same guarantees on the performance.

Instead of using $3\binom{n}{3}$ triangle inequalities to ensure that $d_{i,j}$ is a metric (or pseudo-metric as defined later), we show that only a compact subset of inequalities, so-called *sparse metric*, are sufficient to obtain the same fractional optimal solution.

A function d is a pseudo-metric if $d(i, j) = d_{i,j}$ satisfy the following conditions:

- 1) $d(i, j) \geq 0$ (non-negativity)
- 2) $d(i, i) = 0$ (and possibly $d(i, j) = 0$ for some distinct values $i \neq j$)
- 3) $d(i, j) = d(j, i)$ (symmetry)
- 4) $d(i, j) \leq d(i, k) + d(k, j)$ (transitivity).

It is clear that d is an feasible solution of $\text{LP}_{\text{complete}}$ if and only if d is a pseudo-metric within the interval $[0, 1]$.

Our new linear programming with the *Sparse Metric* tech-

nique, denoted by $\text{IP}_{\text{sparse}}$, is as follows:

$$\text{maximize} \quad -\frac{1}{2m} \sum_{i,j} B_{i,j} d_{i,j} \quad (7)$$

$$\text{subject to} \quad \begin{aligned} d_{i,k} + d_{k,j} &\geq d_{i,j} & k \in N(i, j) \\ d_{i,j} &\in \{0, 1\}, \end{aligned} \quad (8)$$

The objective can be simplified to $-\frac{1}{2m} \sum_{i,j} B_{i,j} d_{i,j}$ since $\sum_{i,j} B_{i,j} = 0$. Let $N(i)$ and $N(j)$ denote the set of neighbors of i and j , respectively. The set $N(i, j)$ is defined as the union of neighbors of i and j

$$N(i, j) = N(i) \cup N(j) - \{i, j\}$$

Therefore, the total number of constraints in the formula is upper bounded by

$$\sum_{i < j} d_i + d_j = (n-1) \sum_{i=1}^n d_i = O(mn)$$

When the considered network is sparse, which is often true for complex networks, our new formulation substantially reduces time and memory requirements. For most real-world network instances, where $n \approx m$, the number of constraints is effectively reduced from $\theta(n^3)$ to $O(n^2)$. If we consider the time to solve linear programming to be cubic time the number of constraints, the total time complexity for sparse networks improves to $O(n^6)$ instead of $O(n^9)$ as in the original approach. In practice, LPs can be solved quite efficiently. We mention the increase of the size of the largest solved instance of traveling salesman problem from 49 cities in 1954 [18] to 85,9000 cities in 2009 [19] as an example of rapid development of mathematical programming solvers and computer powers.

Again, we can obtain the relaxation of $\text{IP}_{\text{sparse}}$, described in (7) to (9), by replacing the constraints $d_{i,j} \in \{0, 1\}$ by $d_{i,j} \in [0, 1]$. We shall refer to this relaxation of $\text{IP}_{\text{sparse}}$ as $\text{LP}_{\text{sparse}}$. The fractional optimal solution of this relaxation can also be rounded and tuned with the same algorithms in the previous subsection.

C. Correctness and Performance Guarantees

In order to achieve the same guarantees provided by solving $\text{LP}_{\text{complete}}$, we show the equivalence of the sparse formulation and the complete formulation:

- $\text{IP}_{\text{sparse}}$ and $\text{IP}_{\text{complete}}$ share the same set of optimal integral solutions (Theorem 1).
- The optimal fractional solutions of $\text{LP}_{\text{sparse}}$ and $\text{LP}_{\text{complete}}$ have same objective values (Theorem 2) i.e. they provide the same upper bound on the maximum possible modularity.

Hence, solving $\text{LP}_{\text{sparse}}$ indeed gives us an optimal solution of $\text{LP}_{\text{complete}}$, while doing so significantly reduces the time and memory requirements.

Theorem 1: Two integer programmings $\text{IP}_{\text{sparse}}$ and $\text{IP}_{\text{complete}}$ share the same set of optimal solutions.

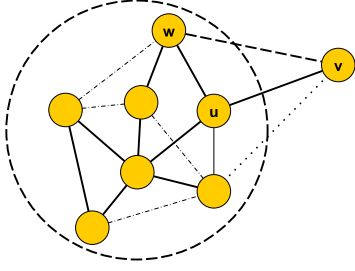


Fig. 1: Clique expanding process.

Proof: We need to show that every optimal solution of $\text{IP}_{\text{complete}}$ is also a solution of $\text{IP}_{\text{sparse}}$ and vice versa.

In one direction, since the constraints in $\text{IP}_{\text{sparse}}$ is a subset of constraints in $\text{IP}_{\text{complete}}$, every optimal solution of $\text{IP}_{\text{complete}}$ will also be a solution of $\text{IP}_{\text{sparse}}$.

In the other direction, let $d_{i,j}$ be an optimal integral solution of $\text{IP}_{\text{sparse}}$. We shall prove that $d_{i,j}$ must be a pseudo-metric that implies $d_{i,j}$ is also a feasible solution of $\text{IP}_{\text{complete}}$.

For convenience, we assume that the original graph $G = (V, E)$ has no isolate vertices that were known to have no affects on modularity maximization [10]. Construct a graph $G_d = (V, E_d)$ in which there is an edge (i, j) for every $d_{i,j} = 0$. Let $C_d = \{C_d^1, C_d^2, \dots, C_d^l\}$ be the set of connected components in G_d , where C_d^t represents the set of vertices in t th connected components.

Proposition 1: Every connected component C_d^i induces a connected subgraph in $G = (V, E)$.

Proof: We prove by contradiction. Assume that the connected component C_d^t does not induce a connected subgraph in G . Hence, we can partition C_d^t into two subsets S and T so that there are no edges between S and T in G .

Construct a new solution d' from d by setting $d'_{i,j} = 1$ for all pairs $(i, j) \in P(S, T)$, the set of pairs with one end point in S and one endpoint in T . Since, $A_{i,j} = 0 \forall (i, j) \in P(S, T)$, we have $B_{i,j} = A_{i,j} - \frac{d_{i,j}}{2m} < 0 \forall (i, j) \in P(S, T)$. Hence, setting $d'_{i,j} = 1 \forall (i, j) \in P(S, T)$ can only increase the objective value. In fact, doing so will strictly increase the objective. There must be at least one pair $(i, j) \in P(S, T)$ with $d_{i,j} = 0$, or else C_d^i is not a connected component in G_d .

It is not hard to verify that $d'_{i,j}$ satisfy all constraints of $\text{IP}_{\text{sparse}}$ since those triangle inequalities must involve at least one edge in the original graph G , while S and T are disconnected sets in G .

Thus, we have derived from an optimal solution a new feasible solution with higher objective (contradiction). ■

The rest is to prove that for each connected component C_d^t of G_d , if $i, j \in C_d^t$ then the distance $d_{i,j} = 0$. We prove by repeatedly applying a “clique expanding” process. At each step, every pair of vertices in the clique are proven to have distance 0. Then, we expand the clique, adding one more adjacent vertex to the clique and prove that the new clique also has vertices of distance zero from each other (see Fig. 1).

Initial step. We first prove there is an edge $(i, j) \in E$ of the original graph G satisfying $d_{i,j} = 0$. We shall choose that edge as our initial clique of size 2. Assume no such edge exists, all

pairs $d_{i,j} = 0$ within C_d^t have $A_{i,j} = 0$ and $B_{i,j} < 0$. Thus, again we can increase the distance of all pairs with $d_{i,j} = 0$ to 1 without violating any constraints, while increasing the objective value (contradiction). Therefore, we can always find an edge that belongs to both G and G_d .

Expanding steps. Denote our clique by K_t . If $K_t = C_d^t$, then we can complete the proof for C_d^t . Otherwise, there is a vertex $u \in K_t$ and a vertex $v \in C_d^t - K_t$, so that (u, v) is an edge in both G and G_d ($d_{u,v} = 0$). The existence of such an edge (u, v) can be proven by contradiction (Assume not, then increase distance of all pairs in $P(K_t, C_d^t - K_t)$ from 0 to 1 to increase the objective value while not violating any constraints.). Then, for each vertex $w \in K_t - \{u\}$, the constraint $d_{w,u} + d_{u,v} \geq d_{w,v}$ is in $\text{IP}_{\text{sparse}}$ and $d_{w,u} = 0$ from the property of K_t . It follows that $d_{w,v} = 0$ for all $w \in K_t$. By adding v to K_t we increase the size of the clique, while ensuring the zero-distance property.

Since the size of C_d^t is at most n , the expanding process will finally terminate with $K_t = C_d^t$. ■

Theorem 2: $\text{LP}_{\text{sparse}}$ and $\text{LP}_{\text{complete}}$ share the same set of fractional optimal solutions.

Proof: We need to show that every fractional optimal solution of $\text{LP}_{\text{complete}}$ is also a fractional solution of $\text{LP}_{\text{sparse}}$ and vice versa. Since the integrality constraints have been dropped in both LP relaxations, we need a different approach to the proof in Theorem 1.

One direction is easy, every fractional optimal solution of $\text{LP}_{\text{complete}}$ is also a fractional solution of $\text{LP}_{\text{sparse}}$.

For the other direction, let $d_{i,j}$ be a fractional optimal solution of $\text{LP}_{\text{sparse}}$, we shall prove that $d_{i,j}$ is also a feasible solution of $\text{LP}_{\text{complete}}$.

Associate a weight $w_{i,j} = d_{i,j}$ for each edge $(i, j) \in E$ (other edges are assigned weights ∞). Let $d'_{i,j}$ be the distance between two nodes (i, j) with the new edge weights. We have

- 1) $d'_{i,j} \geq d_{i,j}$ for all i, j and $d'_{i,j} = d_{i,j} \forall (i, j) \in E$.
- 2) $d'_{i,j} = \min_{k=1}^n \{d'_{i,k} + d'_{k,j}\}$. Hence, $d'_{i,j}$ is a pseudo-metric.

The first statement can be shown by applying the triangle inequalities in $\text{LP}_{\text{sparse}}$. Since, $d'_{i,j}$ be the shortest distance between i and j in G , there is a path $u_0 = i, u_1, \dots, u_l = j$ with the length $d'_{i,j} = d_{u_0,u_1} + d_{u_1,u_2} + \dots + d_{u_{l-1},u_l}$. Since (u_{k-1}, u_k) are edges in G for all $k = 1..l$, we can apply triangle inequalities iteratively

$$\begin{aligned} d_{i,j} &\leq d_{u_0,u_1} + d_{u_1,u_l} \leq d_{u_0,u_1} + d_{u_1,u_2} + d_{u_2,u_l} \\ &\leq \dots \leq d_{u_0,u_1} + d_{u_1,u_2} + \dots + d_{u_{l-1},u_l} = d'_{i,j} \end{aligned} \quad (10)$$

If $(i, j) \in E$, we have $d'_{i,j} \leq d_{i,j}$. Hence, $d'_{i,j} = d_{i,j} \forall (i, j) \in E$. The second statement comes from the definition of $d'_{i,j}$.

Notice that $d'_{i,j}$ may be no longer upper bounded by one. Therefore, we define $d^*_{i,j} = \min\{d'_{i,j}, 1\}$. We also have

$$d^*_{i,j} \geq d_{i,j} \forall i, j \text{ and } d^*_{i,j} = d_{i,j} \forall (i, j) \in E.$$

And more importantly, d^* is also a pseudo-metric. Since $d^*_{i,k} + d^*_{k,j} \geq \min\{d'_{i,k} + d'_{k,j}, 1\} \geq \min\{d'_{i,j}, 1\} = d^*_{i,j}$.

Now, if $d_{i,j} = d_{i,j}^*$ for all i, j , then d satisfies all triangle inequalities in $\text{LP}_{\text{complete}}$ and we yield the proof.

Otherwise, assume that $d_{i,j} < d_{i,j}^*$ for some pair (i, j) . We show that d^* is a feasible solution of $\text{LP}_{\text{sparse}}$ with greater objective value that contradicts the hypothesis that d is an optimal solution.

Since for all edges $(i, j) \notin E$, $d_{i,j} = d_{i,j}^*$, and for pairs $(i, j) \notin E$, $B_{i,j} < 0$ and $d_{i,j}^* \geq d_{i,j}$, we have $\sum_{i,j} d_{i,j}^* > \sum_{i,j} d_{i,j}$ (contradiction). ■

IV. APPROXIMATION ALGORITHMS FOR MAXIMIZING MODULARITY IN POWER-LAW NETWORKS

This section presents approximation algorithms for the modularity maximization problem in power-law networks. A factor ρ approximation algorithm for a maximization problem, find in polynomial-time a solution with the value no less than ρ times the value of an optimal solution. Approximation algorithms are being used for problems where exact polynomial-time algorithms are too expensive and in many cases, they can yield valuable insights to the problem.

We make a detour to focus on the problem of modularity maximization in division of the network into just two communities. The maximum modularity value of the division into two communities are shown to “close” to the best possible modularity. Thus, an approximation algorithm for the division into two communities problem also yields an approximation algorithm for the modularity maximization problem.

A. Division into k Communities

Let Q_k be the maximal modularity obtained by a division of the network into **exact** k communities. We also denote $Q_k^+ = \max_{i=1}^k Q_i$ and $Q_{\text{opt}} = Q_n^+$, the best possible modularity over all possible divisions. Let δ^{opt} be a community structure with the maximum modularity Q_{opt} .

Proposition 2: $Q_1 = 0$ and $Q_n = -\frac{\sum_i d_i^2}{4m^2}$.

Lemma 1:

$$Q_k^+ \geq (1 - \frac{1}{k})Q_{\text{opt}}$$

Proof: If δ^{opt} has at most k communities, then we have $Q_k^+ = Q_{\text{opt}}$. Otherwise δ^{opt} has more than k communities.

We can rewrite the modularity as

$$Q_{\text{opt}} = \frac{1}{2m} \sum_{\delta_{ij}^{\text{opt}}=1} B_{ij}$$

Construct a k -division of the network by randomly assigning communities in δ^{opt} into one of k new “super” communities. Let δ^k denote the obtained partitioning. If $\delta_{ij}^{\text{opt}} = 1$, then $\delta_{ij}^k = 1$ i.e. all within intra-communities pairs remain within new “super” communities. All pairs (i, j) with $\delta_{ij}^{\text{opt}} = 0$ (inter-community pairs) become intra-communities pairs with probability $1/k$. Hence, the contribution of a pair (i, j) with $\delta_{ij}^{\text{opt}} = 0$ to the expected modularity is $\frac{1}{k}B_{ij}$. Hence, the expected modularity of the k -division by randomly grouping

communities will be

$$\begin{aligned} Q_E &= \frac{1}{2m} \left(\sum_{\delta_{ij}^{\text{opt}}=1} B_{i,j} + \frac{1}{k} \sum_{\delta_{ij}^{\text{opt}}=0} B_{i,j} \right) \\ &= \frac{1}{2m} \left(1 - \frac{1}{k} \right) \sum_{\delta_{ij}^{\text{opt}}=1} B_{i,j} = \left(1 - \frac{1}{k} \right) Q_{\text{opt}} \end{aligned}$$

In the second step, we have used the equality $\sum_{i,j} B_{i,j} = 0$ or equivalently $\sum_{\delta_{ij}^{\text{opt}}=1} B_{i,j} = -\sum_{\delta_{ij}^{\text{opt}}=0} B_{i,j}$. Therefore, we have $Q_k^+ \geq Q_E = (1 - \frac{1}{k})Q_{\text{opt}}$. ■

It follows from Lemma 1 that an approximation algorithm with a factor ρ for maximizing Q_2 will also be an approximation with a factor 2ρ to the modularity maximization problem.

For a division of the network into two groups define

$$x_i = \begin{cases} 1, & \text{if } i \text{ belong to community 1} \\ -1, & \text{if } i \text{ belong to community 2.} \end{cases}$$

We can write the modularity for the division into two communities as

$$Q = \frac{1}{4m} \sum_{i,j} B_{i,j}(x_i x_j + 1) = \frac{1}{4m} \sum_{i,j} B_{i,j} x_i x_j = \frac{1}{4m} x^T B x$$

Hence, the division into two communities is a special case of the maximizing quadratic program problem i.e. the problem of finding a vector $x \in \{-1, 1\}^n$ such that $x^T B x$ is maximized. The following results was due to M. Charikar et al. [15] and Nesterov et al. [20].

Theorem 3: [15] Given an arbitrary matrix A , whose diagonal elements are nonnegative, the problem of finding $x \in \{-1, 1\}^n$ such that $x^T B x$ is maximized can be approximated within $O(\log n)$. In case B is positive definite, the ratio can be improved to $\frac{\pi}{2}$ [20].

Unfortunately, the matrix B is not positive definite. Even worse, the main diagonal contains all negative entries as the i th entry is $-\frac{d_i^2}{4m^2}$. Hence, we cannot directly apply above results for the division into two communities problem.

B. Power-law Networks

Complex networks including social, biological, and technology networks display a non-trivial topological feature: their degree sequences can be well-approximated by a power-law distribution [5]. At the same time they exhibit modular property i.e. the existence of naturally division into communities. We establish the connection between the power-law degree distribution property and the modular property, stating that whenever a network have power-law degree distribution, there is presence of communities in the network with a significant modularity.

We use the well-known $P(\alpha, \beta)$ model by F. Chung and L. Lu [21] for power-law networks in which there are y vertices of degree x , where x and y satisfy $\log y = \alpha - \beta \log x$. In other words,

$$|\{v : d(v) = x\}| = y = \frac{e^\alpha}{x^\beta}$$

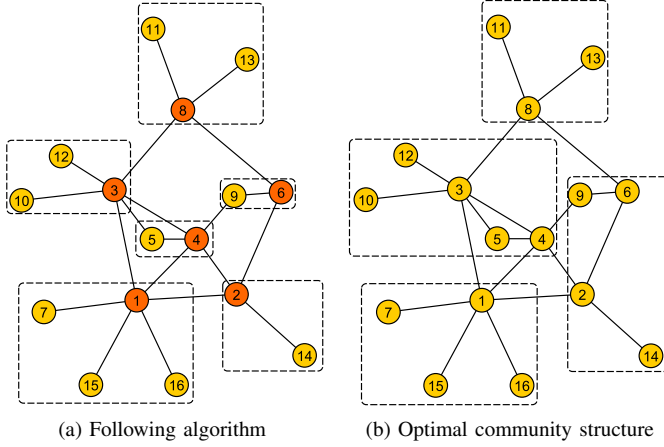


Fig. 2: On the left, a community structure found by Following Algorithm in Theorem 4 when $d_0 = 2$. Each rounded square represents a community, and followers are in the darker color. The modularity is 0.325 i.e. 87% of the optimal modularity, 0.374. On the right, the optimal community structure found by solving IP_{sparse} .

Basically, α is the logarithm of the size of the graph ($n = e^\alpha$) and β is the log-log growth rate of the graph. While the scale of the network depends on α , β decides the connection pattern and many other important characterizations of the network. Different networks at different scales with same β often exhibit same characteristics. For instance, the larger β , the sparser and the more “power-law” the network is. Hence, β is regarded as a constant in $P(\alpha, \beta)$ model.

In $P(\alpha, \beta)$ model, the maximum degree in a $P(\alpha, \beta)$ graph is $e^{\frac{\alpha}{\beta}}$. The number of vertices and edges are

$$n = \sum_{x=1}^{e^{\frac{\alpha}{\beta}}} \frac{e^\alpha}{x^\beta} \approx \begin{cases} \zeta(\beta)e^\alpha & \text{if } \beta > 1 \\ \alpha e^\alpha & \text{if } \beta = 1 \\ \frac{e^{\frac{\alpha}{\beta}}}{1-\beta} & \text{if } \beta < 1 \end{cases},$$

$$m = \frac{1}{2} \sum_{x=1}^{e^{\frac{\alpha}{\beta}}} x \frac{e^\alpha}{x^\beta} \approx \begin{cases} \frac{1}{2} \zeta(\beta-1)e^\alpha & \text{if } \beta > 2 \\ \frac{1}{4} \alpha e^\alpha & \text{if } \beta = 2 \\ \frac{1}{2} \frac{e^{\frac{2\alpha}{\beta}}}{2-\beta} & \text{if } \beta < 2 \end{cases} \quad (11)$$

where $\zeta(\beta) = \sum_{i=1}^{\infty} \frac{1}{i^\beta}$ is the Riemann Zeta function. Without affecting the conclusions, we will simply use real number instead of rounding down to integers. The error terms can be easily bounded and are sufficiently small in our proofs.

Most real-world networks have the log-log growth rate β between 2 and 3. For examples, scientific collaboration networks with $2.1 < \beta < 2.45$ [22], Word Wide Web with β for in-degree and out-degree of 2.1 and 2.45, respectively [23]; Internet at router and intra-domain level with $\beta = 2.48$ and so on. No power-law networks with $\beta < 1$ have been observed. One of the reason is that when $\beta < 1$, the number of edges $m = \Omega(n^2)$ i.e. the network is not “scale-free”.

Theorem 4: There is an $O(\log n)$ approximation algorithm for the modularity maximization problem in power-law networks with the log-log growth rate $\beta > 1$. If $\beta > 2$, the

TABLE I: Order and size of network instances

Problem ID	Name	Nodes n	Edges m
1	Zachary’s karate club	34	78
2	Dolphin’s social network	62	159
3	Les Miserables	77	254
4	Books about US politics	105	441
5	American College Football	115	613
6	US Airport 97	332	2126
7	Electronic Circuit (s838)	512	819
8	Scientific Collaboration	1589	2742

problem can be approximated within a constant approximation factor $2\zeta(\beta-1)$, where $\zeta(x) = \sum_{i=1}^{\infty} \frac{1}{i^x}$ is the Riemann Zeta function.

Proof: From Lemma (1) with $k = 2$, we have $\frac{1}{2}Q_{\text{opt}} \leq Q_2^+$. Hence, it is sufficient to approximate Q_2^+ within a factor of $O(\log n)$.

We have

$$Q_2^+ = \frac{1}{4m} \max_{x \in \{-1,1\}^n} x^T B x$$

$$= \frac{1}{4m} \max_{x \in \{-1,1\}^n} x^T B_0 x - \sum_{i=1}^n \frac{d_i^2}{8m^2}, \quad (12)$$

where B_0 is obtained by replacing the diagonal of B with zeros.

Let $D = \sum_{i=1}^n \frac{d_i^2}{8m^2}$, the second term in equation (12). We can approximate

$$\text{OPT}_0 = \max_{x \in \{-1,1\}^n} x^T B_0 x = Q_2^+ + D$$

within a factor of $O(\log n)$ by the method in Theorem 3. That means we can find a division of the network into two communities with the modularity is at least

$$\frac{c}{\log n} \text{OPT}_0 - D = \frac{c}{\log n} (Q_2^+ + D) - D$$

$$\geq \frac{c}{\log n} Q_2^+ - D \geq \frac{c}{2 \log n} Q_{\text{opt}} - D$$

where c is an independent constant.

If we can show that $D = o\left(\frac{1}{\log n} \text{OPT}_0\right)$, then we can approximate the maximum modularity within a factor $O(\log n)$. This is equivalent to

$$\lim_{n \rightarrow \infty} \frac{Q_{\text{opt}}}{D \log n} = \infty \text{ or } \lim_{\alpha \rightarrow \infty} \frac{Q_{\text{opt}}}{D \log n} = \infty \quad (13)$$

To show (13), we present a linear-time algorithm, called *Following*, to find a community structure \mathcal{L} with a lower bound on the modularity. An illustration example for the algorithm is shown in Fig. 2a.

Following Algorithm (Parameter $d_0 \in \mathbb{N}^+$)

- i. Start with all nodes unlabeled
- ii. Sort nodes in *non-decreasing order of degree*
- iii. For each unlabeled node v with $d_v \leq d_0$, find a neighbor u that is not a follower; set v to follow u i.e. label v “follower” and u “followee”. If many such u exist, select the one with the minimum degree.
- iv. Label all unlabeled nodes “followee”.
- v. Put each followee and its followers into a community.

Despite that higher values of d_0 possibly lead to better approximation ratios, it is sufficient for our proof to consider only the case $d_0 = 1$. That means all leaf nodes will attach to (follow) their neighbors. Assume that for a graph $G = (V, E)$, vertices in V are numbered so that leaf nodes will have higher numbering than non-leaf nodes i.e. $V = \{v_1, v_2, \dots, v_t, \underbrace{v_{t+1}, \dots, v_n}_{\text{leaf nodes}}\}$ in which t is the number

of non-leaf nodes. For a node $v_i, i = 1 \dots t$, let $l_i \leq d_i$ be the number of leaves attached to v_i . There will be t communities associated with v_1, v_2, \dots, v_t , respectively.

Since there are e^α vertices of degree one, there are at least $\frac{1}{2}e^\alpha$ edges inside considered communities. Hence,

$$\begin{aligned} Q(\mathcal{L}) &= \frac{e^\alpha}{2m} - \sum_{i=1}^t \frac{(d_i + l_i)^2}{4m^2} \geq \frac{e^\alpha}{2m} - \sum_{i=1}^n \frac{4d_i^2}{4m^2} \\ &= \frac{e^\alpha}{2m} - 8D \end{aligned} \quad (14)$$

Since $Q_{\text{opt}} \geq Q(\mathcal{L})$, instead of showing (13), we can show

$$\lim_{\alpha \rightarrow \infty} \frac{Q(\mathcal{L})}{D \log n} = \infty \Leftrightarrow \lim_{\alpha \rightarrow \infty} \frac{e^\alpha/2m}{D \log n} = \infty$$

From the power-law degree distribution in (11):

$$D = \sum_{x=1}^{\frac{\alpha}{\beta}} \frac{e^\alpha}{x^\beta} \frac{x^2}{8m^2} = \frac{e^\alpha}{8m^2} \sum_{x=1}^{\frac{\alpha}{\beta}} x^{2-\beta} \quad (15)$$

Consider all three cases of β :

Case $\beta > 2$: Since $x^{2-\beta} < 1$, from equation (11) we have

$$\begin{aligned} Q(\mathcal{L}) &\geq \frac{e^\alpha}{2m} - 8D \geq \frac{1}{\zeta(\beta-1)} - \frac{4e^{\frac{\alpha}{\beta}}}{\zeta(\beta-1)^2 e^\alpha} \\ &\geq \frac{1}{2\zeta(\beta-1)} \end{aligned} \quad (16)$$

Since $Q_{\text{opt}} \leq 1$, community structure \mathcal{L} approximate the optimum solutions within a constant factor $2\zeta(\beta-1)$.

Case $\beta = 2$: We have $\log n < 2\alpha$. Hence,

$$D \log n \leq \frac{2e^\alpha}{\alpha^2 e^{2\alpha}} \left(\sum_{x=1}^{\frac{\alpha}{\beta}} 1 \right) 2\alpha = \frac{4e^{\alpha/\beta}}{\alpha e^\alpha}$$

Thus,

$$\lim_{\alpha \rightarrow \infty} \frac{e^\alpha/2m}{D \log n} \geq \lim_{\alpha \rightarrow \infty} \frac{e^\alpha}{2e^{\alpha/\beta}} = \infty$$

Hence, the modularity maximization problem can be approximated within a factor $O(\log n)$ in this case.

Case $2 > \beta > 1$:

$$\begin{aligned} D \log n &\leq \frac{e^\alpha}{8m^2} e^{\frac{\alpha}{\beta}(3-\beta)} \sum_{x=1}^{\frac{\alpha}{\beta}} \left(\frac{x}{e^{\frac{\alpha}{\beta}}} \right)^{2-\beta} \frac{1}{e^{\frac{\alpha}{\beta}}} 2\alpha \\ &\leq \frac{2\alpha e^\alpha}{(2-\beta)^2 e^{\frac{4\alpha}{\beta}}} e^{\frac{\alpha}{\beta}(3-\beta)} \int_0^1 x^{2-\beta} dx \\ &\leq \frac{(2-\beta)^2}{e^{\frac{\alpha}{\beta}}} \frac{\alpha}{3-\beta} \end{aligned}$$

Therefore,

$$\begin{aligned} \lim_{\alpha \rightarrow \infty} \frac{e^\alpha/2m}{D \log n} &\geq \lim_{\alpha \rightarrow \infty} \frac{e^\alpha}{2e^{\frac{2\alpha}{2-\beta}}} \frac{(3-\beta)e^{\alpha/\beta}}{\alpha(2-\beta)^2} \\ &\geq \lim_{\alpha \rightarrow \infty} \frac{3-\beta}{\alpha(2-\beta)} e^{\alpha(1-\beta^{-1})} = \infty \end{aligned}$$

Hence, the theorem follows. \blacksquare

TABLE II: The modularity obtained by previous published methods GN [5], EIG [10], VP [13], LP_{complete} [13], our *sparse metric* approach LP_{sparse} and the optimal modularity values OPT [14]. The optimal modularity for network 8 (as a whole) has not been known before; we compute it by solving our our IP_{sparse} within only 15 seconds.

ID	n	GN	EIG	VP	LP _{complete}	LP _{sparse}	OPT
1	34	0.401	0.419	0.420	0.420	0.420	0.420
2	62	0.520	-	0.526	0.529	0.529	0.529
3	77	0.540	-	0.560	0.560	0.529	0.529
4	105	-	0.526	0.527	0.527	0.529	0.529
5	115	0.601	-	0.605	0.605	0.605	0.605
6	332	-	-	-	-	0.368	0.368
7	512	-	-	-	-	0.819	0.819
8	1589	-	-	-	-	0.955	0.955

V. COMPUTATIONAL EXPERIMENTS

We present experimental results for our linear programming rounding algorithm in Section III. The LP solver is GUROBI 4.5, running on a PC computer with Intel 2.93 Ghz processor and 12 GB of RAM. We evaluate our algorithm on several standard test cases for community structure identification, consisting of real-world networks. The datasets names together with their sizes are listed in Table I. The largest network consists of 1580 vertices and 2742 edges. All references on datasets can be found in [13] and [14].

TABLE III: Number of constraints in formulations LP_{complete} used in paper [13] (Constraint(C)) and the computational time (in seconds) (Time(C)) versus number of constraints in our *sparse metric* formulation LP_{sparse} (Constraint(S)) and its computational time(Time(S)).

ID	n	Constraint(C)	Constraint(S)	Time(C)	Time(S)
1	34	17,952	1,441	0.21	0.02
2	62	113,460	5,743	3.85	0.11
3	77	219,450	6,415	13.43	0.08
4	105	562,380	30,236	60.40	1.76
5	115	740,715	66,452	106.27	13.98
6	332	18,297,018	226,523	-	197.03
7	512	66,716,160	294,020	-	53.18
8	1589	2,002,263,942	159,423	-	2.94

Since the same rounding procedure are applied on the optimal fractional solutions, both LP_{complete} and LP_{sparse} yield the same modularity values. However, LP_{sparse} can run on much larger network instances. The modularity of the rounding LP algorithms and other published methods are shown in Table II. The rounding LP algorithm can find optimal solutions (or within 0.1% of the optimal solutions) in all cases. The source code for our LP algorithm can be obtained upon request.

Finally, we compare the number of constraints of the LP formulation used in [13] and our new formulation (LP_{sparse}) in Table III. Our new formulation contains substantially less constraints, thus can be solved more effectively. The old LP formulation cannot be solved within the time allowance (10000 seconds) and the memory availability (12 GB) in cases of the network instances 6 to 8. The largest instance of 1589 nodes is solved surprisingly fast, taking under 3 seconds. The reason is due to the presence of leaves (nodes of degree one) and other special motifs that can be efficiently preprocessed with the reduction techniques in [24].

Our new technique substantially reduces the time and memory requirements both theoretically and experimentally without any trade-off on the quality of the solution. The size of solved network instances raises from hundred to several thousand nodes while the running time on the medium-instances are sped up from 10 to 150 times. Thus, the *sparse metric* technique is a suitable choice when the network has a moderate size and a community structure with performance guarantees is desired.

VI. DISCUSSION

We have proposed two algorithms for the modularity maximization problem in complex networks. Our algorithms successfully exploit sparseness and power-degree distribution property found in many complex networks to provide performance guarantees on the solutions. On one hand, the algorithms implied in Theorem 4 are the first approximation algorithms for maximizing modularity, hence, are of theoretical interest. On the other hand, our *sparse metric* approach is an efficient method to find optimal or close to optimal community structure for networks of up to thousand nodes.

Fortunato and Barthélemy [25] have recently shown that in general quality functions of global definitions of community, including modularity, has an intrinsic resolution scale, known as resolution limit. Therefore, they fail to detect communities smaller than a scale, which depends on global attributes of networks such as the total size and the degree of connection among communities. However, resolution limit can be overcome by introducing a scaling parameter $\lambda > 0$ into the original modularity formula as independently proposed by Arenas et al. [26] and R. Lambiotte et al. [27].

$$Q_\lambda(C) = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \lambda \frac{d_i d_j}{2m} \right) \delta_{i,j}$$

Our proposed methods work naturally with this extension with little modification. The only changes in the LP formulations are in the objective coefficients; the modularity matrix B is replaced with a new “multi-scale” modularity matrix B^λ with $B_{i,j}^\lambda = A_{i,j} - \lambda \frac{d_i d_j}{2m}$. The *sparse metric* technique still applies and provides the same guarantees as solving the complete LP formulation. In addition, the constant λ does not affect the *asymptotic* approximation ratios of algorithms in Theorem 4. Our ongoing work is to design an efficient modularity approximation algorithm that both gives a better approximation ratio and perform well in practice.

REFERENCES

- [1] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, 1998.
- [2] A. Barabasi, R. Albert, and H. Jeong, “Scale-free characteristics of random networks: the topology of the world-wide web,” *Physica A*, vol. 281, 2000.
- [3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: simple building blocks of complex networks,” *Science (New York, N.Y.)*, vol. 298, no. 5594, 2002.
- [4] S. Fortunato and C. Castellano, “Community structure in graphs,” *Encyclopedia of Complexity and Systems Science*, 2008.
- [5] M. Girvan and M. E. Newman, “Community structure in social and biological networks,” *PNAS*, vol. 99, no. 12, 2002.
- [6] W. H. E. Day and H. Edelsbrunner, “Efficient algorithms for agglomerative hierarchical clustering methods,” *Journal of Classification*, vol. 1, 1984.
- [7] J. Reichardt and S. Bornholdt, “Statistical mechanics of community detection,” *Phys. Rev. E*, vol. 74, 2006.
- [8] A. Gog, D. Dumitrescu, and B. Hirsbrunner, “Community detection in complex networks using collaborative evolutionary algorithms,” in *Advances in Artificial Life*, ser. LNCS. Springer Berlin / Heidelberg, 2007, vol. 4648.
- [9] J. Duch and A. Arenas, “Community detection in complex networks using extremal optimization,” *Phys. Rev. E*, vol. 72, no. 2, 2005.
- [10] M. E. J. Newman, “Modularity and community structure in networks,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, 2006.
- [11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, 2008.
- [12] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, “On modularity clustering,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 2, 2008.
- [13] G. Agarwal and D. Kempe, “Modularity-maximizing graph communities via mathematical programming,” *Eur. Phys. J. B*, vol. 66, no. 3, 2008.
- [14] D. Aloise, S. Cafieri, G. Caporossi, P. Hansen, S. Perron, and L. Liberti, “Column generation algorithms for exact modularity maximization in networks,” *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, vol. 82, 2010.
- [15] M. Charikar and A. Wirth, “Maximizing quadratic programs: Extending grothendieck’s inequality,” *FOCS*, 2004.
- [16] N. Bansal, A. Blum, and S. Chawla, “Correlation clustering,” in *Machine Learning*, 2002.
- [17] B. W. Kemighan and S. Lin, “An efficient heuristic procedure for partitioning graphs,” *Journal of Classification*, 1970.
- [18] G. Dantzig, R. Fulkerson, and S. Johnson, “Solution of a large-scale traveling-salesman problem,” *Operations Research*, vol. 2, 1954.
- [19] D. L. Applegate, R. E. Bixby, V. Chvatal, W. Cook, D. G. Espinoza, M. Goycoolea, and K. Helsgaun, “Certification of an optimal tsp tour through 85,900 cities,” *Operations Research Letters*, vol. 37, no. 1, 2009.
- [20] Y. Nesterov, “Semidefinite relaxation and nonconvex quadratic optimization,” CORE Discussion Papers 1997044, 1997.
- [21] W. Aiello, F. Chung, and L. Lu, “A random graph model for massive graphs,” in *STOC '00*. New York, NY, USA: ACM, 2000.
- [22] A. L. Barabasi, H. Jeong, Z. Nda, E. Ravasz, A. Schubert, and T. Vicsek, “Evolution of the social network of scientific collaborations,” *Physica A: Statistical Mechanics and its Applications*, vol. 311, 2002.
- [23] R. Albert, H. Jeong, and A. Barabasi, “Error and attack tolerance of complex networks,” *Nature*, vol. 406, 2000.
- [24] D. J. F. A. G. S. Arenas, A., “Size reduction of complex networks preserving modularity,” *New J. Phys.*, vol. 9, 2007.
- [25] S. Fortunato and M. Barthélemy, “Resolution limit in community detection,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, 2007.
- [26] A. Arenas, A. Fernandez, and S. Gomez, “Analysis of the structure of complex networks at different resolution levels,” *New J. Phys.*, vol. 10, 2008. [Online]. Available: doi:10.1088/1367-2630/10/5/053039
- [27] R. Lambiotte, J. C. Delvenne, and M. Barahona, “Laplacian dynamics and multiscale modular structure in networks,” *arXiv*, vol. 812, 2008.